DEPARTMENT OF COMMERCE
SINCLAIR WEEKS, Secretary

WEATHER BUREAU
F. W. REICHELDERFER, Chief

MONTHLY WEATHER REVIEW

Editor, JAMES E. CASKEY, JR.

Volume 81
Number 11

NOVEMBER 1953

Closed January 15, 1954
Issued February 15, 1954

THE USE OF PROBABILITY STATEMENTS IN EXTENDED FORECASTING

WALTER G. LEIGHT [1]
Extended Forecast Section, U. S. Weather Bureau, Washington, D. C.
[Manuscript received July 17, 1953; minor revisions received October 20, 1953]

ABSTRACT

The need for qualification of weather forecasts has long been recognized. The diverse uses of predictions and variations in the forecaster's ability to prognosticate weather events correctly have suggested to previous writers that expressions of confidence or other means of weighting can serve the public to advantage. A particular method of expressing and scoring 5-day temperature forecasts in terms of probability statements has been tried experimentally by the U. S. Weather Bureau's Extended Forecast Section and found to provide valuable information in addition to that normally furnished. Furthermore, it is demonstrated that the participants in the experiment possess definite skill in selecting probability statements, and that relative verification results are comparable to those obtained by scoring the conventional forecasts.

INTRODUCTION

Since weather forecasting is not an exact science, an individual prediction can at best be the forecaster's opinion of the weather events most likely to follow the antecedent conditions. Current practice in both short-range and extended forecasting involves the expression of this opinion in terms of categorical statements, the deficiencies of which have been discussed by investigators dating back at least to 1906 (Cooke [4]) and, more recently by Thompson [16]. It is the purpose of this paper to explore the need for a different form of expression, namely probability statements; to investigate the mechanics of expression for extended forecasts; and to determine the practicality of instituting a program of this type.

THE NEED FOR PROBABILITY INFORMATION

Although conventional forecasts are of obvious value to the public, they suffer from three major defects independent of their accuracy. In the first place, the forecast user is often interested in the likelihood of occurrence of a weather event which is not considered by the forecaster as "most likely" among the possible eventualities and hence, conceivably, is left unmentioned in the ordinary categorical forecast (Hallenbeck [9], Price [14]).

A second defect is inherent in the fact that the same categorical forecast may be issued by the same forecaster on two occasions of considerably different circumstances. This problem was recognized almost 50 years ago by Cooke [4] who stated: "It seems to me that the condition of confidence or otherwise forms a very important part of the prediction, and ought to find expression." This position was stated even more strongly in a report of a Special Committee on Forecast Verification, appointed by the Chief of the U. S. Weather Bureau in 1939, and quoted by Brier [1]: ". . . a mere forecast of an event without some qualification when the forecaster knows that there is considerable uncertainty would be withholding from the public information to which it is entitled." In the past, however, qualification has generally been expressed in the form of vague "weasel words". Only in the light of recent technical advances has it become possible to review the problem and evaluate the forecaster's capabilities along these lines.

The practical significance of furnishing all possible forecast information to the consumer derives from the fact that forecast value varies with the forecast user (Brier in "Panel Discussion on Forecast Verification" [13], and Brier and Allen [3]). Similarly, Thompson [16]

[1] Present affiliation: Operations Evaluation Group, Navy Department, Washington, D. C.

concludes that issuing categorical weather predictions to individuals or agencies charged with making operating decisions ". . . in many cases places an appreciable handicap on the most effective use of the prediction."

The third shortcoming of categorical weather forecasts stems from the verification of predictions. In attempts to improve performance scores, forecasters often resort to "hedging" which, according to Brier [2], "may lead the forecaster to forecast something other than that which he thinks will occur, for it is often easier to analyze the effect of different possible forecasts on the verification score than it is to analyze the situation."

EXPRESSION AND USE OF CONFIDENCE STATEMENTS

Many investigators have advanced proposals to include or substitute statements of the probability of occurrence of the predicted weather events in order to circumvent one or more of the difficulties described above. Cooke [5], for example, suggested that a single figure be appended to each forecast to indicate the probability subjectively assigned by the forecaster. Thompson [16] also recommends that the consumer's problems of basing operational decisions can be relieved by " providing an estimate of the probability of occurrence of critical weather and permitting the user to make his own operating decision." This procedure is also supported by Brier and Allen [3] who advocate a scheme of probability forecasting and verification "that cannot influence the forecaster in any undesirable way. . . . (The) forecaster is encouraged to minimize [i. e., in their context, optimize] his score by getting the forecasts exactly right and stating a probability of unity. If he cannot forecast perfectly, he is encouraged to state unbiased estimates of the probability of each possible event. . . . The user can then interpret the information in terms of his operation."

An ideal scheme of expression of forecasts has been aptly described by Gringorten [7]: "if the forecaster could state the percent likelihood of each of all possible mutually exclusive events, he would be giving a complete answer as to his evaluation of the weather prospects." Of course, it is immediately self-evident that the crux of the matter lies in the forecaster's ability to analyze the situation and correctly assess likelihood of occurrence. Although a question has also been raised as to the ability of the general public to comprehend probability statements, it is usually recognized that most laymen readily understand the "odds" quoted for sporting events. Of even greater importance, however, is the fact that a complete statement of the probabilities of occurrence of all possible weather events automatically includes an indication of what the forecaster considers to be the most likely, that is to say, the conventional categorical forecast. Furthermore, with reference to the experiment to be described below, the 5-day temperature forecasts prepared by the Extended Forecast Section are not transmitted to the general public directly, but rather to District Forecast

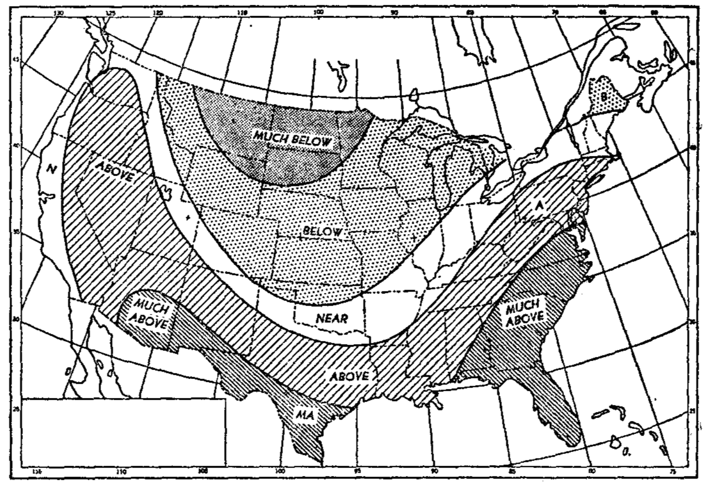

FIGURE 1.—Five-day temperature forecast (schematic) in terms of anomaly categories.

Centers where the recipients are well equipped to interpret probability statements.

From the forecaster's standpoint, it is also worth mentioning that the use of complete forecast information—including estimates of events not expected to occur—provides a better measure of his predictive ability if we agree with a definition of skill advanced by Gringorten [8]: "the forecaster's ability to analyze and classify the antecedent weather so that, within one class, the probability of a subsequent event is increased above, or decreased below, the relative frequency of that event in all weather situations (. . . the 'climatic frequency')."

EVALUATION OF PROBABILITY

The methods proposed for stating confidence in a categorical forecast or complete probability information over the full range of possible occurrences have varied from entirely subjective assessments, as discussed by Cooke [5] and Williams [17], to wholly objective techniques, as recommended by Dickey [6], Gringorten [7], and Thompson [15]. These writers, however, have concerned themselves almost exclusively with the problem of expressing the forecast for a single meteorological variable at a single locale, although extending the scope of the predictions increases only the procedural difficulties and not the basic principles involved.

At the Weather Bureau's Extended Forecast Section, circulation, temperature, and precipitation prognoses are prepared twice weekly and transmitted to regional meteorological centers where they may be modified on the basis of latest developments and knowledge of local characteristics before dissemination to the public. The temperature predictions, with which we are concerned in this paper, are made in terms of five categories of departure from normal: *near* normal, *above* and *below* normal, *much above* and *much below* normal (Namias [11, 12]). These categories are defined on the basis of records at individual stations for each month of the year in such a way that

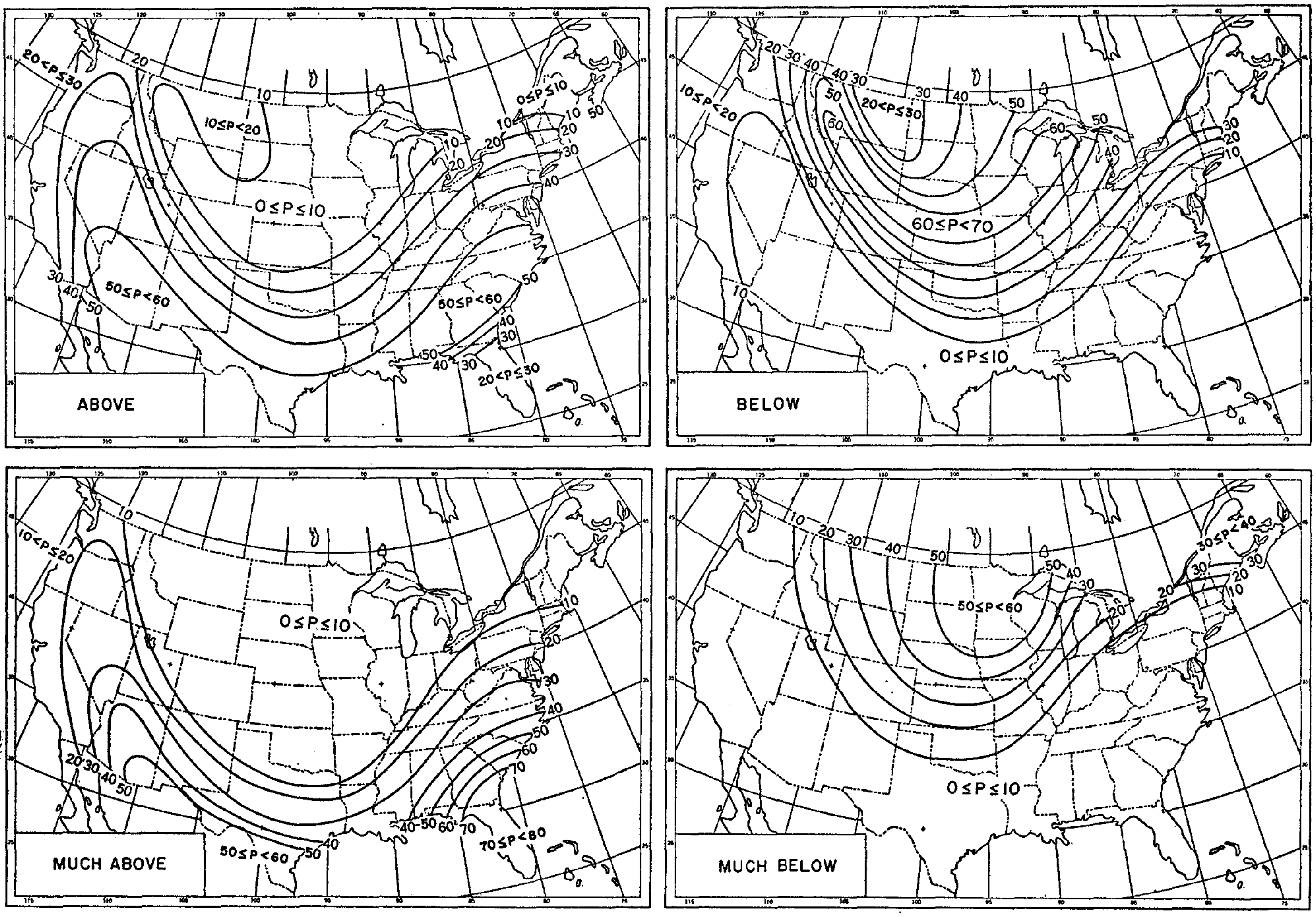

FIGURE 2.—Predicted probabilities of occurrence of separate categories for same period as in figure 1. Probability of *near* normal is calculated as 1.00 minus (sum of probabilities of other four classes).

each of the first three categories just mentioned occurred one-fourth of the 5-day periods within the record studied, while the latter two each occurred one-eighth of the time. These definitions not only simplify the expression of the conventional temperature forecast, as illustrated in figure 1, but also provide a convenient standard of comparison, based on climatology and chance, for measuring forecast skill (see Namias [12], and Brier and Allen [3]).

Objective methods devised for preparing 5-day temperature anomaly forecasts (Martin and Leight [10]), although adaptable, are not at present in suitable form for assessing the likelihood of occurrence of each of the five possible categories. It was therefore decided to experiment with probability forecasting by allowing the official forecaster to stipulate subjective estimates.

For five categories the probability forecast for the United States area can best be presented by a series of charts, as shown in figure 2, instead of the single chart shown in figure 1. A separate map shows for each category the geographical distribution of expected likelihood of occurrence. When prepared in this fashion, care must be taken to insure that at every point the sum of the probabilities of the five classes must be exactly unity. In the illustration shown, the chart for *near* normal is omitted; the probability of *near* normal at any point can be readily computed by subtracting from 1.00 the sum of the probabilities of the other four classes.

In order to simplify the procedure, the forecasters were provided with a code sheet, shown in figure 3, containing 22 selected statements of the probability of occurrence of each of the five temperature anomaly categories shown at the tops of the columns. The first digit of the two-figure code number indicates the "most likely" category, i. e., the conventional categorical forecast, while the second digit increases, essentially, with increasing likelihood of the warmer classes. Obviously, these statements represent a very small sample of the infinite variety of statements which a forecaster might choose subjectively, but it was felt that increasing the number of lines in the table would only serve to complicate the evaluation process and might, at the same time, furnish a spurious indication of the forecaster's capability in delineating the expected probabilities of occurrence. The specific choices were made only after deliberations by the personnel

PROBABILITY FORECASTS

CLASS FORECAST	CODE NUMBER	PROBABILITY MB	B	N	A	MA
1 Much Below	11	.90	.10	.00	.00	.00
	12	.75	.25	.00	.00	.00
	13	.50	.50	.00	.00	.00
	14	.40	.30	.20	.10	.00
	4 YR. AVG.	.22	.44	.20	.11	.03
2 Below	21	.50	.50	.00	.00	.00
	22	.30	.50	.20	.00	.00
	23	.20	.60	.20	.00	.00
	24	.33	.34	.33	.00	.00
	25	.20	.40	.30	.10	.00
	4 YR. AVG.	.12	.34	.26	.21	.07
3 Near Normal	31	.10	.25	.50	.15	.00
	32	.00	.33	.34	.33	.00
	33	.00	.20	.60	.20	.00
	34	.00	.15	.50	.25	.10
	4 YR. AVG.	.11	.29	.28	.26	.06
4 Above	41	.00	.10	.30	.40	.20
	42	.00	.00	.33	.34	.33
	43	.00	.00	.20	.60	.20
	44	.00	.00	.20	.50	.30
	45	.00	.00	.00	.50	.50
	4 YR. AVG.	.03	.16	.24	.36	.21
5 Much Above	51	.00	.10	.20	.30	.40
	52	.00	.00	.00	.50	.50
	53	.00	.00	.00	.25	.75
	54	.00	.00	.00	.10	.90
	4 YR. AVG.	.02	.08	.14	.34	.42
CLIMATOLOGY		.125	.25	.25	.25	.125

FIGURE 3.—Code sheet for subjective selection of probability statements.

charged with making the official five-day forecasts and, as might be expected, experimentation has led to many suggestions for improvements.

For guidance purposes, the code sheet also included the lines labelled "4-yr avg." and "climatology." The former figures represent the fraction of time during the 4-year period immediately preceding the experiment that each of the classes was observed when a specific one was categorically predicted. It may be of interest to compare these figures with those in table 1 which was similarly computed for the 11-year period from October 1940 through September 1951. The high degree of similarity between the 2 sets of figures demonstrates that real skill is involved in making the conventional categorical forecasts. The actual values show the extent of that skill and, at the same time, emphasize the advantages of providing the consumer with estimates of the likelihood of occurrence of the anomaly classes not considered "most likely."

TABLE 1.—*Frequency of occurrence of temperature anomalies, by class, during the period October 1940 through September 1951 according to anomaly class predicted*

Predicted class	Observed class 1-MB	2-B	3-N	4-A	5-MA	Total
1-MB	0.246	0.387	0.212	0.122	0.033	1.00
2-B	.116	.338	.276	.213	.057	1.00
3-N	.053	.247	.295	.305	.100	1.00
4-A	.025	.149	.241	.386	.199	1.00
5-MA	.013	.074	.144	.375	.394	1.00

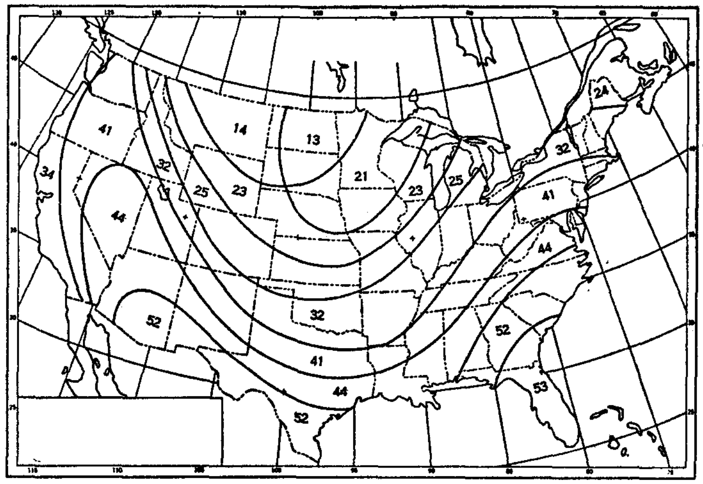

FIGURE 4.—Temperature forecast for same period as in figures 1 and 2 in terms of coded statements shown in figure 3.

The "climatology" line on the code sheet (fig. 3) simply expresses the probability of occurrence of each of the classes in accordance with the definitions.

Using this code sheet the five official forecasters participating in the experiment prepared probability forecasts, as shown in figure 4 for the same situation illustrated in figures 1 and 2, for a total of 50 forecast periods. The prognoses were prepared under the operational pressures of the regular forecast routine on Monday and Thursday evenings, immediately after completion of the categorical forecast (as in figure 1).

METHOD OF VERIFICATION

As proposed by Brier [2], the probability forecasts were verified pointwise by summing over the five categories the squares of the differences between predicted and observed probabilities,

$$S_1=\sum_1^5 (P_f-P_o)^2 \tag{1}$$

where P_f is the probability forecast and P_o is that observed (unity for occurrence and zero for non-occurrence). For each forecast S_1, the sum of the squared errors of the probability estimates, was computed for each of 100 points, converted to a score,

$$S^*=1-\tfrac{1}{2}\, S_1 \tag{2}$$

in order to define a score which increased with increasing success, and a final score,

$$S=\sum_1^{100} S^* \tag{3}$$

was determined. This arbitrarily transformed [2] score,

[2] By reversing the transformations one may easily reconvert the results presented below to the mean squared errors which are perhaps more significant from a statistical point of view.

S, obtained by summing scores for 100 individual points, had the conventional range from 0, for the worst possible forecasts, to 100, for perfect predictions. In actuality, by virtue of the coded statements used, it was impossible to obtain either extreme score.

It is worth noting from the above that for each point and for each forecast period there were, essentially, five estimates of the probability of occurrence of the five temperature anomaly classes. Since there were 50 forecast periods, and since the verification was conducted for 100 points evenly distributed over the United States, 5,000 forecasts were involved. It is to be noted, of course, that, for several reasons, these are not at all independent.

It is usually considered desirable to compare forecast scores with those obtainable by other methods or blind devices, including persistence, chance, etc. (Brier and Allen [3]). Accordingly, for this test scores were also computed on the following bases:

(1) Assuming that the forecasters possessed no skill, and hence for each point-forecast the best probability estimate was that furnished by theoretical climatology (last line of the code sheet in fig. 3) by virtue of the definition of classes.

(2) Assuming no forecast skill and using the observed frequency distribution of temperature anomaly classes over the preceding four years as the climatological standard of comparison. This differs from theoretical climatology in that the 4-year period was anomalously warm; its use is predicated on a continuation of the skewed distribution of temperature classes.

(3) Accepting the categorical forecast and using as probability statement the frequency distribution achieved by forecast class during the previous four years (the lines labelled "4-yr. avg.").

(4) Accepting the categorical prediction and using as probability statement that of "least confidence" (code number 14, 25, 32, 41, or 51).

From the above-mentioned scores and those for the subjective probability predictions various analyses were made of the relative ability of forecasters in estimating likelihood of occurrence (by checking observed frequencies against forecast probabilities), etc.

DISCUSSION OF VERIFICATION SCORES

A summary of the verification scores is shown by the bar graphs in figure 5. The results obtained by the five forecasters participating (labelled "A" through "E"), by the group as a whole ("ALL") and by the "blind" forecasts have been compared to the "expected" and are expressed as skill scores in accordance with the formula:

$$\text{Skill Score} = \frac{S-E}{T-E} \times 100 \tag{4}$$

where S is the score obtained, E the "expected", and T the maximum score attainable.

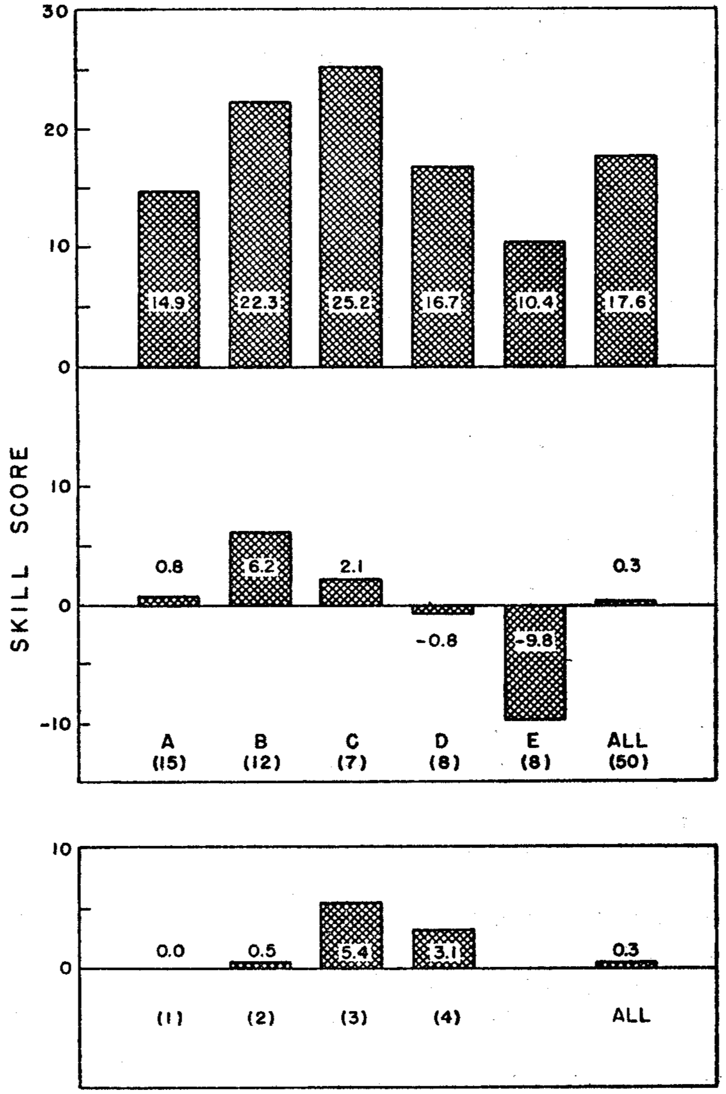

FIGURE 5.—Skill scores obtained by individual forecasters (A, B, C, D, and E) and group as whole (ALL) for categorical forecasts (upper) and for probability statements (middle); numbers in parentheses indicate number of predictions per forecaster. Skill scores obtained by "blind" controls, as enumerated in text (1, 2, 3, and 4) and by group of forecasters as a whole (ALL) are shown at bottom.

The skill scores for the conventional categorical forecasts are shown in the upper part of figure 5. For each forecast period the skill score is regularly computed by taking as S the number of points of 100 distributed over the United States (T) for which the category was predicted correctly. The "expected" is calculated as one-eighth of the number of points for which either *much above* or *much below* is observed [3] plus one-fourth of the remaining points of the total 100. This is the number correct one would most likely have obtained by drawing forecasts randomly from a population distributed in proportion to the defined category distribution.

The forecasts made in terms of probability statements

[3] The choice of "expected" is arbitrary, and is often made in terms of persistence, the distribution of the predicted marginal totals, or a combination of observed and predicted totals for each category; each method has desirable and undesirable characteristics. The "expected" based on the distribution of observed categories, used in routine verification and, therefore, in this study, is independent of manipulation by the forecaster.

were verified by using equations (1), (2), and (3), and score S, provided by equation (3), was then used as S in equation (4). Instead of an "expected" based on chance, the value of E was based on climatology in the following way: the theoretical climatology (last line in figure 3) was used as an unvarying probability statement for each point forecast, thus giving another set of scores from equations (1), (2), and (3). The value of T is again 100 for each forecast for the United States.

The skill scores from this comparison are presented in the middle section of figure 5. Although the numerical values of skill scores for the two kinds of forecast are not comparable, it is interesting that the relative rankings among forecasters are not appreciably different. Two of the skill scores in the second section of the figure are actually negative, demonstrating that, with the particular scoring system used, it is difficult to show superiority to climatology (see remarks by Mook in Panel Dicussion [13]). This difficulty may be ascribed mainly to two factors: first, the nature of the scoring which, on an individual forecast, favors the statement which "spreads" the probabilities among those classes which do not occur; and, secondly, the forecaster's overconfidence in his categorical forecasts (see below) which also includes inexperience in the method. This seems to be substantiated by the third section of figure 5 which includes the results obtained by the "blind" forecasts enumerated in the previous section. The bars labelled "3" and "4" are based on the forecaster's selection of "most likely" category, but favor that class less than the subjective statements. At the same time, the probability that the category will not verify exactly is, in those two cases, distributed among three or four of the remaining categories (the other statements on the code sheet do not permit such distribution).

One might conclude from the success of the "4-yr avg." forecasts (item 3) that statistics of past performance are perhaps the best source of estimates of probability. Obviously, they provide an excellent first approximation to estimates of likelihood of occurrence—which does not imply perfect forecasting. On the contrary, it is recognized that our forecasting ability is far from perfect, and probability statements, such as the automatically-translated "4-yr avg." estimates, supply valuable additional information to the recipient of the categorical forecasts.

It remains to be demonstrated whether or not the forecaster can subjectively improve on the statistics of past performance. The fact that the "least confidence" statements (item 4), while not quite as good as the "4-yr avg." lines, are measurably better than the climatological (items 1 and 2) and subjective (ALL) estimates buttresses the contention that the scoring is aided by the "spreading" of likelihood of occurrence among all categories. (It should be stressed that over a sufficiently long period the best scores will derive from probability estimates which are in accord with the observed frequency distributions. The success of the "4-yr avg." follows from its derivation from relatively long-term records.) If any improvements are to be made subjectively, they must depend on the forecasters' ability to distinguish between situations, i. e., capability in correctly delineating degree of confidence.

ABILITY TO EXPRESS CONFIDENCE

As stated above, each of the 5000 point-forecasts made during the experiment represented five sub-forecasts of the probability of occurrence of each of the temperature anomaly categories ranging from .00 to .90. To measure the forecasters' skill in selecting statements, the predictions were classified into five groups as shown in table 2 with, as an example, the pertinent code numbers (from fig. 3) for the *much above* normal predictions. Analyses were then performed for each participating forecaster and each predicted category to relate confidence, as expressed in the code numbers, to observed frequency of occurrence. The results are summarized in figures 6 and 7.

The data in figure 6 have been plotted at the mid-points of the range groups just defined against the percentage of the point forecasts that the particular category was

TABLE 2.—*Probability forecast groupings for analysis of forecasters' capabilities*

Group	Probability range	Applicable code numbers for forecasts of *much above* normal
I	.00 and .10	11 through 34.
II	.15 through .25	41 and 43.
III	.30 through .40	42, 44, and 51.
IV	.50 and .60	45 and 52.
V	.75 and .90	53 and 54.

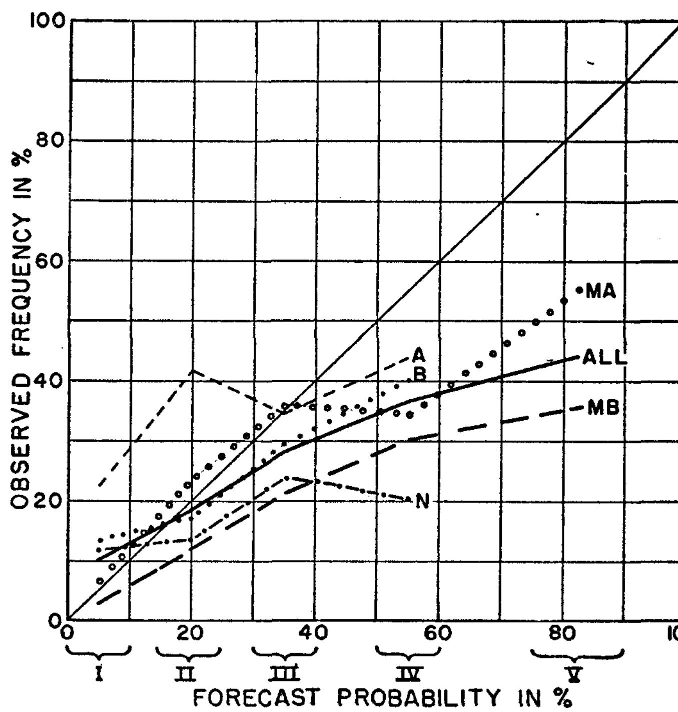

FIGURE 6.—Relationship between forecast probability, plotted at midpoints of range groups indicated by Roman numerals, and observed frequency of occurrence of individual temperature anomaly categories and all categories combined. Perfect probability estimates represented by 45° line.

observed. This has been done for each category for all forecasters combined and, in addition, a composite curve has been drawn for all categories combined. The six curves can also be compared with the 45° line which indicates perfect ability in sorting weather situations into groups for which the forecaster knows the percentage frequency of occurrence.

Inspection of the curves in figure 6 reveals, as might be expected from the figures in table 1, that forecasting the *near* normal category is most difficult. This has also been found to be true in a qualitative fashion by the forecasters themselves, probably because *near* normal is usually observed in strips separating large areas of greater anomaly. Figure 6 also shows marked overconfidence in the higher ranges of estimate for the extreme categories and for *above* normal in the lower ranges. However, it is encouraging to note the distinct upward slope of the curves, showing definite increase of observed frequency with greater confidence. (It might be noted that the lines would be horizontal if there were absolutely no skill and would slope downward to the right with negative skill.)

Similar conclusions can be drawn from figure 7 which combines all temperature categories to show the relationship of predicted vs. observed probability (frequency) by forecaster. As in figure 6, the 45° line of "perfect estimate" and the composite for all forecasters and categories (representing 25,000 point-category forecasts) are included. This figure also contains a horizontal dashed line representing the distribution that might be expected from a random sampling of a population distributed according to the definition of classes.

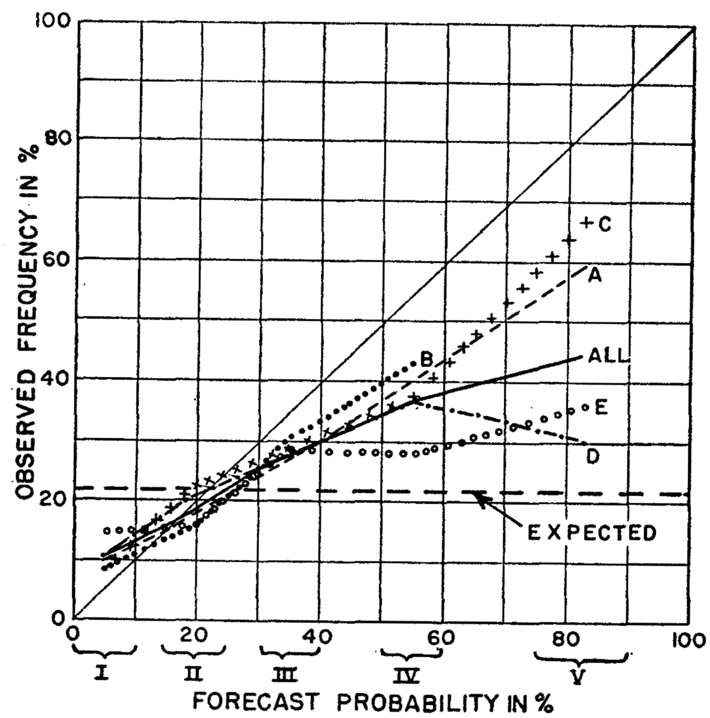

FIGURE 7.—Relationship between forecast probability, plotted at mid-points of range groups indicated by Roman numerals, and observed frequency of occurrence for all temperature categories combined by individual forecaster and group as whole. Perfect estimates shown by 45° line and chance estimates shown by dashed horizontal line.

FREQUENCY OF USE OF STATEMENTS

It should be borne in mind that the overconfidence portrayed in figures 6 and 7 is primarily associated with the statements involving extremely great likelihood of occurrence of the forecast categories, and that in the middle and lower ranges the probability estimates are quite good. It is therefore of interest to determine the relative frequency of use of the five ranges of prediction; this is shown in figure 8 for all forecasters and all categories combined. On this chart there are also two broken bars to indicate the distribution theoretically expected from consistent use of "climatology".

It can be seen from figure 8 that the high ranges of probability were used rather infrequently, albeit overconfidently, during the experiment. There is a relative maximum of use of the middle ranges, for which the best results were obtained, but another maximum exists for Group I: .00 through .10. This can be compensated for by providing better selection of probability statements. As suggested above, this probably would also lead to a bettering of the verification scores.

CONCLUSIONS

In earlier experiments Hallenbeck [9] obtained an almost-straight-line relationship between his probability forecasts and the observed frequencies of occurrence, and Williams' [17] results indicated "that the forecasters at Salt Lake City, for the period under study, on the average, were quite apt to know whether or not their forecasts would verify." On the basis of these experiences and the project herein described it is concluded that forecasters possess the required capability of distinguishing between forecasts which are more or less likely to verify. Accordingly, one is led to the conviction that the verification scores achieved by the "4-yr avg." can be materially

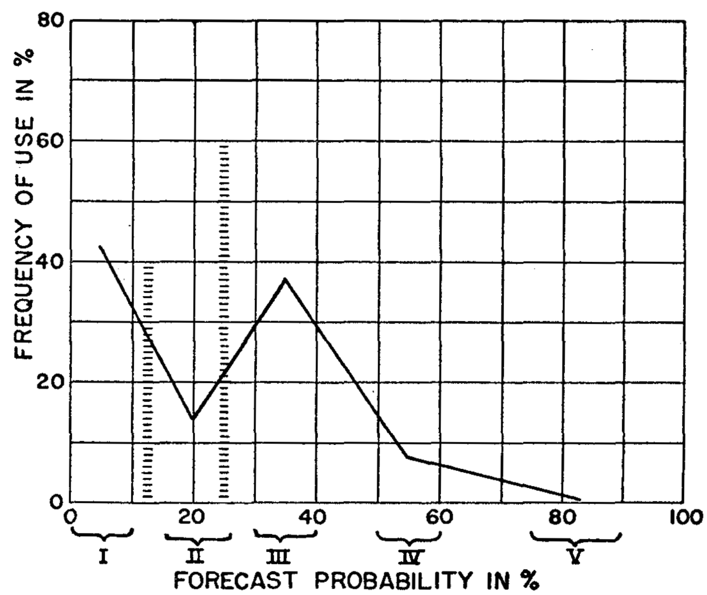

FIGURE 8.—Frequency of use of probability statements as grouped in table 2. Vertical bars indicate forecasting by climatology only.

improved by judicious use of coded statements more carefully contrived.

The improvement of the statements presented for the forecasters' consideration might best be achieved by using the "4-yr avg." lines of figure 3 (or the 11-yr avg. data from table 1) as a guide and introducing into the tabulation estimates involving more reasonable "spreading" of the probability of nonoccurrence of the category considered "most likely". It would also be feasible to include "blank" code numbers (such as 19, 29, etc.) for which the forecaster might specify probability ranges required for a particular forecast problem. In regions dominated by a quasi-stationary front during a forecast period, for example, the forecaster might well wish to assign equal probabilities of cold and warm anomalies, with a minimum probability of *near* normal.

Since the conduct of this experiment demonstrated that the additional labor involved in the preparation of the probability estimates was not great, this scheme is operationally practical at the transmitting end. In fact, the participating forecasters have estimated that the availability of better statements would reduce the time required to an insignificant amount, relatively speaking.

From the consumer's point of view, it seems incontrovertible that the extra information provided by probability statements (Thompson [16])—or even by such statistics as presented in table 1—is of great value. Since it has been demonstrated that the forecasters can, with fair degree of skill, assess confidence in the predictions, this information should be made available to the forecast user.

In conclusion, it is felt that some system of expressing forecasts in terms of probability is practical and extremely desirable. Such estimates can be prepared without undue extra effort on the part of the forecasters and with a fair amount of skill. Furthermore, the form of expression discussed here is susceptible to scoring with results comparable to those obtained by verification of conventional categorical forecasts, while at the same time including automatically indication of the category considered "most likely" as well as additional valuable information.

REFERENCES

1. G. W. Brier, "Verification of a Forecaster's Confidence and the Use of Probability Statements in Weather Forecasting," U. S. Weather Bureau *Research Paper No. 16*, 1944, 10 pp.
2. G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, vol. 78, No. 1, January 1950, pp. 1–3.
3. G. W. Brier and R. A. Allen, "Verification of Weather Forecasts," *Compendium of Meteorology*, American Meteorological Society, Boston, 1951, pp. 841–848.
4. W. E. Cooke, "Forecasts and Verifications in Western Australia," *Monthly Weather Review*, vol. 34, No. 1, January 1906, pp. 23–24.
5. W. E. Cooke, "Weighting Forecasts," *Monthly Weather Review*, vol. 34, No. 6, June 1906, pp. 274–275.
6. W. W. Dickey, "Estimating the Probability of a Large Fall in Temperature at Washington, D. C.," *Monthly Weather Review*, vol. 77, No. 3, March 1949, pp. 67–78.
7. I. I. Gringorten, "Forecasting by Statistical Inferences," *Journal of Meteorology*, vol. 7, No. 6, December 1950, pp. 388–394.
8. I. I. Gringorten, "The Verification and Scoring of Weather Forecasts," *American Statistical Association Journal*, vol. 46, No. 255, Sept. 1951, pp. 279–296.
9. C. Hallenbeck, "Forecasting Precipitation in Percentages of Probability," *Monthly Weather Review*, vol. 48, No. 11, November 1920, pp. 645–647.
10. D. E. Martin and W. G. Leight, "Objective Temperature Estimates from Mean Circulation Patterns," *Monthly Weather Review*, vol. 77, No. 10, October 1949, pp. 275–283.
11. J. Namias, *Extended Forecasting by Mean Circulation Methods*, U. S. Weather Bureau, 1947, 89 pp.
12. J. Namias, "Thirty-Day Forecasting—A Review of a Ten-Year Experiment," *Meteorological Monographs*, vol. 2, No. 6, American Meteorological Society, 1953, 83 pp.
13. "Panel Discussion on Forecast Verification," *Bulletin of the American Meteorological Society*, vol. 33, No. 7, September 1952, pp. 274–278.
14. S. Price, "Thunderstorm Today?—Try a Probability Forecast," *Weatherwise*, vol. 2. No. 3, June 1949, pp. 61–63.
15. J. C. Thompson, "A Numerical Method for Forecasting Rainfall in the Los Angeles Area," *Monthly Weather Review*, vol. 78, No. 7, July 1950, pp. 113–124.
16. J. C. Thompson, "On the Operational Deficiencies in Categorical Weather Forecasts," *Bulletin of the American Meteorological Society*, vol. 33, No. 6, June 1952, pp. 223–226.
17. Philip Williams, Jr., "The Use of Confidence Factors in Weather Forecasting," *Bulletin of the American Meteorological Society*, vol. 32, No. 8, October 1951, pp. 279–281.